University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2015

Spatio-phylogenetic multispecies distribution models

Kaldhusdal, Arne ; Brandl, Roland ; Müller, Jörg ; Möst, Lisa ; Hothorn, Torsten

Abstract: 1. Ecologists increasingly consider phylogenetic relatedness in both community composition and spatial arrangements in communities. Here we considered both the phylogenetic correlation between multiple species and the spatial correlation induced by unobserved spatial heterogeneity on multiple plots. For this analysis, we introduced phylogenetic spatial generalised linear mixed models (PSGLMMs), which are an extension of phylogenetic generalised linear mixed models (PGLMMs). 2. We used the framework of generalised linear array models to simultaneously model species and plot dimension. Such models have the potential to explain the correlation of the phylogenetic relationship of the observed species and of the spatial proximity of the plots, or both. We proposed model selection strategies based on proper scores and empirically evaluated them in a case study using bird count data. In our analysis, we focused on two special cases: the community composition model and the environmental sensitivity model. 3. Our simulation study indicated that it might be difficult to correctly identify phylogenetic signals when the phylogenetic correlation is rather low and when studying presence-absence or count data of rare or pervasive species.

DOI: https://doi.org/10.1111/2041-210X.12318

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-115524
Journal Article
Accepted Version

Originally published at:
Kaldhusdal, Arne; Brandl, Roland; Müller, Jörg; Möst, Lisa; Hothorn, Torsten (2015). Spatio-phylogenetic multispecies distribution models. Methods in Ecology and Evolution, 6(2):187-197.
DOI: https://doi.org/10.1111/2041-210X.12318

Spatio-Phylogenetic Multi-Species Distribution Models

Arne Kaldhusdal[1], Roland Brandl[2], Jörg Müller[3,4], Lisa Möst[1] and Torsten Hothorn[5]

[1]Institut für Statistik, LMU München, Germany

[2]Fachbereich Biologie, Philipps-Universität Marburg, Germany

[3]Sachgebiet Forschung und Dokumentation, Nationalparkverwaltung Bayerischer Wald, Grafenau, Germany

[4]Lehrstuhl für Terrestrische Ökologie, Technische Universität München, Germany

[5]Institut für Epidemiologie, Biostatistik und Prävention, Universität Zürich, Switzerland

Number of words abstract: 174

Number of words text: 6639

Number of figures: 3

Number of tables: 5

Number of references: 31

Number of online appendices: 2

Abstract

Ecologists increasingly consider phylogenetic relatedness in both community composition and spatial arrangements in communities. Here we considered both the phylogenetic correlation between multiple species and the spatial correlation induced by unobserved spatial heterogeneity on multiple plots. For this analysis, we introduced phylogenetic spatial generalised linear mixed models (PSGLMMs), which are an extension of phylogenetic generalised linear mixed models (PGLMMs). We used the framework of generalised linear array models to simultaneously model species and plot dimension. Such models have the potential to explain the correlation of the phylogenetic relationship of the observed species and of the spatial proximity of the plots, or both. We proposed model selection strategies based on proper scores and empirically evaluated them in a case study using bird count data. In our analysis, we focused on two special cases: the community composition model and the environmental sensitivity model. Our simulation study indicated that it might be difficult to correctly identify phylogenetic signals when the phylogenetic correlation is rather low and when studying presence-absence or count data of rare or pervasive species.

Keywords environmental gradient; generalised linear array models, null-model; phylogenetic community structure; phylogenetic signal; spatial corre-

lation

1 Introduction

With the increasing availability of phylogenetic information, two lines of ecological and biogeographical inquiries have arisen: those leading to an understanding of adaptation (e.g. Hansen & Orzack, 2005) and those leading to an understanding of processes that influence community composition and the establishment of regional biotas (e.g. Lovette & Hochachka, 2006; Stephens & Wiens, 2009). If we consider the evolution of species, e.g. as a Brownian motion process (Felsenstein, 1985) in which genetic changes along an evolutionary trajectory are random and generally small, the degree of divergence between a pair of species should be proportional to the time since they diverged through speciation. In such cases, a phylogenetic signal is said to be present between the species for the trait under study (Blomberg & Garland, 2002; Losos, 2008; Revell et al., 2008); for a clear description of the difference between phylogenetic signal and phylogenetic niche conservatism see Losos (2008). As traits mediate ecological processes, it follows that phylogenetic relatedness within and between communities would allow these ecological processes to be inferred. For this task, metrics have been developed to explore whether phylogenetically related species co-occur more often (e.g. by

environmental filtering) or less often (e.g. through competitive interactions) than expected by chance (e.g. Webb et al., 2002; Vamosi et al., 2009). A growing number of studies using metrics, such as the mean phylogenetic distance between species or individuals, in combination with certain null-models have shown that ecological communities are indeed phylogenetically structured (reviewed, e.g. in Vamosi et al., 2009).

However, metrics like the mean phylogenetic distance between species or individuals have a number of drawbacks (Ives & Helmus, 2011), such as uncertainties of selecting the most appropriate and effective metric and null-model for testing the hypothesis under consideration, low statistical power, and the lack of possibilities for predictions. Therefore, Ives & Helmus (2011) proposed model-based statistics in the form of phylogenetic generalised linear mixed models (PGLMMs) to analyse the assembly of communities along environmental gradients. The flexibility of the model allows one to design tests for phylogenetic signals in occurrence data, and these tests have a higher statistical power than tests using metrics and null-models. Furthermore, these models provide possibilities to test for a phylogenetic signal in the sensitivity to an environmental factor. Finally, PGLMMs can even consider the case in which different species show similar reactions to environmental factors, where, after consideration of this attraction, related species tend not to co-occur (Ives & Helmus, 2011).

In our study, we embedded generalised linear mixed models for multiple species observed on multiple plots into the generalised linear array model (GLAM) framework (Currie et al., 2006). We show that PGLMMs as suggested by Ives & Helmus (2011) can be understood as a special case of GLAMs. Within the class of GLAMs, we extended PGLMMs to models that take into account both the phylogenetic correlation between species and the spatial correlation between plots. Within this new class of phylogenetic and spatial generalised linear mixed models (PSGLMMs), the correlation between the observations might be due to the phylogenetic correlation of the species, the spatial correlation of the plots, or both. Thus, efficient model-selection procedures are needed to differentiate between these three situations in a way driven by data, although careful interpretation is mandatory because the best fitting model might not describe the underlying process best. We suggest the application of cross-validated proper scores, such as the Dawid-Sebastiani or the Brier score, for model selection. On a more technical level, we applied standard software for generalised linear mixed model estimation (the R package `lme4`, Bates et al., 2013; R Core Team, 2013) for investigating the empirical performance of the model selection procedures. Special emphasis was given to the development of tests for a phylogenetic signal in the co-occurrence pattern of species (Model I in Ives & Helmus, 2011) as well as tests for a phylogenetic signal in the sensitivity of species to envi-

ronmental factors (Model II in Ives & Helmus, 2011) for spatially correlated observations.

2 Methods

Our phylogenetic and spatial generalised linear mixed models (PSGLMMs) are GLMMs for which the random effects correlation structure is considered known and based on the phylogeny of the species considered. As a result, in addition to the fixed effects parameters, only a single variance parameter had to be estimated for each random effect modelled. The random effects variance parameter estimates were then used to decide whether a phylogenetic signal was present in the data. The model accepts any correlation structure derived from the phylogeny ranging from a correlation matrix based on the Brownian motion model with one free parameter to more complex models with more than one free parameter, like the accelerated (decelerated) model suggested by Blomberg et al. (2003).

The models may in principle be used for data with responses pertaining to any distribution belonging to the exponential family, but we focused on Gaussian, binary, and Poisson responses in our empirical investigations. For species (spp) $s = 1, \ldots, S$ and plots $p = 1, \ldots, P$, the responses were organised in an $(S \times P)$-matrix $\boldsymbol{Y}$, i.e. in a two-dimensional array. Each column thus

represented the responses in a unique plot and each row represented the responses of a unique species. In the case of binary data, a plot-species datum y_{sp} took the value 1 if species s was present in plot p and took the value 0 if the species was absent. For count data, y_{sp} represented the number of observations made of species s in plot p, and for Gaussian data, any real value attribute/observation for species s in plot p may be used. Repeated measurements are possible, in which case $\boldsymbol{Y}$ would be an array of dimension $(S \times P \times n)$, with n being the number of measurements per species and plot (incomplete designs can be filled with weight zero observations). For the sake of simplicity, we refer to the responses as occurrence data.

We assumed two different sources of variance in the species occurrence patterns. First, a species is more likely to occur in a plot if it is also present in neighbouring plots. This influence was determined by a spatial correlation matrix $\boldsymbol{\Sigma}_{\text{plot}}$, based on some sort of neighbourhood/correlation structure of the plots. Second, a species is more likely to occur in a plot if a closely related species is also present (Ives & Helmus, 2011). This influence was determined by a phylogenetic correlation matrix $\boldsymbol{\Sigma}_{\text{spp}}$. The combination of these two sources of variance will be referred to as spatio-phylogenetic variation.

Two different models were set up. The first model, the community composition model (CCM), was used to test for a phylogenetic signal in the co-occurrence of species, i.e. whether their phylogenetic relationship has an

influence on the composition of communities. The second model, the environmental sensitivity model (ESM), was used to test for a phylogenetic signal in the sensitivity of a species to an environmental factor, i.e. whether phylogenetically closely related species respond to an environmental factor in a similar way. The CCM is thus the spatially enhanced version of Ives & Helmus' (2011) model I, and the ESM corresponds to their model II, but takes the spatial correlation of the plots into account.

Following Ives & Helmus (2011), we modelled phylogenetic attraction, where closely related species are more likely to co-occur, using a correlation matrix derived from a phylogenetic tree. To model phylogenetic repulsion, where closely related species are less likely to co-occur, we used the inverse of that correlation matrix as proposed by Ives & Helmus (2011).

2.1 Community Composition Model

In order to construct the design matrices for the CCM, we defined the species and plot matrices

$$\boldsymbol{S} = \boldsymbol{I}_S \in \mathbb{R}^{S \times S} \quad \text{and} \quad \boldsymbol{P} = \boldsymbol{I}_P \in \mathbb{R}^{P \times P},$$

where $\boldsymbol{I}_S$ is an identity matrix of size S, and $\boldsymbol{I}_P$ is an identity matrix of size P. The corresponding phylogenetic and spatial correlation matrices, assumed to

be known and fixed, are denoted

$$\boldsymbol{\Sigma}_{\text{spp}} \in \mathbb{R}^{S\times S} \quad \text{and} \quad \boldsymbol{\Sigma}_{\text{plot}} \in \mathbb{R}^{P\times P}.$$

In order to account for the overall occurrence of the species, the CCM contained species-specific fixed effects. We accounted for the spatio-phylogenetic variation in the species occurrence by including random intercepts for each species and plot, which were assumed to be correlated in accordance with the phylogenetic and spatial correlation matrices. In the form of an array model (Currie et al., 2006), the CCM in its simplest form is

$$\mathbb{E}[\boldsymbol{Y}] = h[\boldsymbol{S}\boldsymbol{B}\mathbf{1}_P^\top + \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{P}],$$

where $h[\cdot]$ is a response (i.e. inverse link) function suitable for the distribution of $\boldsymbol{Y}$ (e.g., inverse logit, exponential, or identity) applied to each element of its argument. The matrix $\boldsymbol{B} \in \mathbb{R}^{S\times 1}$ contains the species fixed effects coefficients (as we only modelled a fixed intercept, $\boldsymbol{B}$ is effectively a vector $\boldsymbol{\beta}_{\text{spp}} \in \mathbb{R}^S$ for the CCM). Because these are the same in all plots, the species matrix $\boldsymbol{S}$ is implicitly expanded by the vector of P ones, $\mathbf{1}_P$, to form a design matrix. In this simple special case, the linear predictor can be written in the simpler form $\boldsymbol{\beta}_{\text{spp}}\mathbf{1}_P^\top + \boldsymbol{\Gamma}$. The matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{S\times P}$ contains the random

effects coefficients for each species-plot datum. These were assumed to be distributed as

$$\text{vec}(\boldsymbol{\Gamma}) = \boldsymbol{\gamma} \sim \mathcal{N}_{SP}(\mathbf{0}, \sigma^2_\gamma \boldsymbol{\Sigma}_{\text{plot}} \otimes \boldsymbol{\Sigma}_{\text{spp}}),$$

where the "vec" operator performs a column-wise concatenation of $\boldsymbol{\Gamma}$ and $\otimes$ denotes the Kronecker product of two matrices. The spatio-phylogenetic variance parameter $\sigma^2_\gamma = \sigma^2_{\text{plot}}\sigma^2_{\text{spp}}$ is the only unknown random effects variance parameter, and an estimate $\hat{\sigma}^2_\gamma > 0$ would indicate the presence of a spatio-phylogenetic signal.

In order for standard mixed model algorithms to be applicable to the estimation of PSGLMMs, we transformed the random effects so that they were uncorrelated. We thus standardised them to be distributed as

$$\text{vec}(\boldsymbol{\Gamma}) = \boldsymbol{\gamma} \sim \mathcal{N}_{SP}(\mathbf{0}, \sigma^2_\gamma \boldsymbol{I}_{SP})$$

while at the same time penalising the random effects design matrices accordingly (this is a standard procedure in linear mixed models, see Fahrmeir et al., 2013). For this, we first required the Cholesky decomposition of the correlation matrix of the random effects. $\boldsymbol{\Sigma}_{\text{plot}}$ and $\boldsymbol{\Sigma}_{\text{spp}}$ were decomposed separately as

$$\boldsymbol{R}_{\Sigma_{\text{plot}}}\boldsymbol{R}_{\Sigma_{\text{plot}}}^{\top} = \boldsymbol{\Sigma}_{\text{plot}} \quad \text{and} \quad \boldsymbol{R}_{\Sigma_{\text{spp}}}\boldsymbol{R}_{\Sigma_{\text{spp}}}^{\top} = \boldsymbol{\Sigma}_{\text{spp}}.$$

171 As the Cholesky decomposition of a Kronecker product equals the Kronecker
172 product of the Cholesky decomposition of each factor, we obtained the overall
173 decomposition

$$(\boldsymbol{R}_{\Sigma_{\text{plot}}} \otimes \boldsymbol{R}_{\Sigma_{\text{spp}}})(\boldsymbol{R}_{\Sigma_{\text{plot}}} \otimes \boldsymbol{R}_{\Sigma_{\text{spp}}})^{\top} = \boldsymbol{\Sigma}_{\text{plot}} \otimes \boldsymbol{\Sigma}_{\text{spp}}.$$

174 Penalising the design matrix of the random effects by the factor ($\boldsymbol{R}_{\Sigma_{\text{plot}}} \otimes$
175 $\boldsymbol{R}_{\Sigma_{\text{spp}}}$) allowed us to re-express the CCM as

$$\mathbb{E}[\boldsymbol{Y}] = h[\boldsymbol{S}\boldsymbol{B}\boldsymbol{1}_P^{\top} + \tilde{\boldsymbol{S}}\boldsymbol{\Gamma}\tilde{\boldsymbol{P}}], \tag{1}$$

176 using the penalised plot and species matrices

$$\tilde{\boldsymbol{P}} := \boldsymbol{P}\boldsymbol{R}_{\Sigma_{\text{plot}}} \quad \text{and} \quad \tilde{\boldsymbol{S}} := \boldsymbol{S}\boldsymbol{R}_{\Sigma_{\text{spp}}}.$$

177 Hereby, the species-plot specific intercepts $\boldsymbol{\Gamma}$ were implicitly standardised by
178 the Cholesky factor $(\boldsymbol{R}_{\Sigma_{\text{plot}}} \otimes \boldsymbol{R}_{\Sigma_{\text{spp}}})^{-1}$.

2.2 Environmental Sensitivity Model

For the ESM, an environmental factor $\boldsymbol{x} \in \mathbb{R}^P$ was measured in each plot. We added fixed effects $\boldsymbol{\beta}_x \in \mathbb{R}^S$ for each species in addition to the fixed effects $\boldsymbol{\beta}_{\text{spp}}$, and thus assumed that each species also has a base sensitivity to the environmental factor. Because of the statistical difficulty in distinguishing between different phylogenetic effects (Ives & Helmus, 2011), the random intercepts used in the CCM were dropped in favour of random slopes in the environmental factor. This accounted for the spatio-phylogenetic variance associated with the sensitivity of the species to the environmental factor. We thus assumed a linear dependency between the environmental factor and the linear predictor of the ESM. When the plot matrix $\boldsymbol{P}$ is replaced with an environmental factor matrix $\boldsymbol{X} = \text{diag}(\boldsymbol{x}) \in \mathbb{R}^{P \times P}$, a diagonal matrix with the values of $\boldsymbol{x}$ on its diagonal, the ESM is expressed as

$$\mathbb{E}[\boldsymbol{Y}] = h[\boldsymbol{S}\boldsymbol{B}\boldsymbol{D}^\top + \tilde{\boldsymbol{S}}\boldsymbol{\Gamma}\tilde{\boldsymbol{X}}] = h[\boldsymbol{\beta}_{\text{spp}}\boldsymbol{1}_P^\top + \boldsymbol{\beta}_x\boldsymbol{x} + \tilde{\boldsymbol{S}}\boldsymbol{\Gamma}\tilde{\boldsymbol{X}}]. \tag{2}$$

As the ESM contains two fixed effects, $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_{\text{spp}} & \boldsymbol{\beta}_x \end{bmatrix} \in \mathbb{R}^{S \times 2}$ is now a matrix and $\boldsymbol{S}$ is accordingly expanded by the matrix $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{1}_P & \boldsymbol{x} \end{bmatrix} \in \mathbb{R}^{P \times 2}$ to implicitly form a design matrix. We also defined the penalised environmental factor matrix as $\tilde{\boldsymbol{X}} := \boldsymbol{X}\boldsymbol{R}_{\Sigma_{\text{plot}}}$. $\boldsymbol{\Gamma}$ was assumed to be distributed as for the CCM. It is possible to model the sensitivity to several environmental

factors in one model by simply adding one random effect component for each factor. However, this would render the model selection step (Section 2.4) considerably more complex.

2.3 Computational Considerations

For the normal case, parameter estimation in linear mixed models is based on either the maximum likelihood estimator or the restricted maximum likelihood estimator. In a first step, the (restricted) maximum likelihood estimator for σ_γ is computed by maximising the (restricted or penalised) profile log-likelihood. Formulating a PSGLMM as an array model helps to speed-up the evaluation of the linear predictor considerably, however, the optimisation procedure and thus also the estimates are not affected, and one can expect to obtain the exact same estimates through optimisation with or without using the array formulation. The estimation of the variance parameters σ^2_{plot} and σ^2_{spp} in models where the number of random effects equals the number of observations (as in the PSGLMMs studied here) is especially challenging since σ^2_{plot} and σ^2_{spp} and the residual variance σ_ϵ are not well identifiable. In this situation, additional challenges arise and the traditional optimisers cannot be expected to converge well.

In the generalised case, where the conditional density of $\boldsymbol{Y}$ is assumed to

follow a distribution from the one-parameter exponential family, model inference is based on the marginal density obtained by integration with respect to the random effects distribution. This integral is approximated and the quality of the estimation of σ_γ heavily depends on how close the marginal log-likelihood can be resembled. This process is completely independent of the array formulation and, as in the normal case, there will be considerable speed-ups to the evaluation of the approximate marginal log-likelihood but the estimates will be the same. There is no residual variance term in the logistic and Poisson models, so the variance σ^2_γ is identifiable.

In the context of anisotropic smoothing, Rodríguez-Álvarez et al. (2014) recently proposed an algorithm for the estimation of the variance parameters based on GLAMs. Although being much faster than a generalised mixed model, the results were virtually the same. So, overall improvements in convergence and accuracy can not be expected from the application of GLAMs alone.

2.4 Model selection

Even if the spatio-phylogenetic variance parameter σ^2_γ is estimated to be non-zero, we have no guarantee that another correlation structure – even one not accounting for phylogeny – might not be better suited to describe

the variance in the observed data. The true correlation structure of the observations is not known, but four possible interesting scenarios are realistic for our setting: (1) the correlation is only phylogenetic, (2) the correlation is only spatial, (3) the correlation is spatio-phylogenetic, or (4) for Gaussian and Poisson responses, it is possible that the species-plot-specific variance is independent of both space and phylogeny. For binary responses, independent variances are not possible because of the constrained variance structure of binary data. For Gaussian responses in the CCM, the case of independent variance may also be neglected as it would be identical to the estimated residual variance. Selecting among different phylogenetic correlation structures allows an assessment of the Brownian motion assumption.

When testing for a phylogenetic signal in species community structure or in the sensitivity of the species to environmental factors, it is important to account for all reasonable correlation structures in the test. To test which correlation structure is actually the most suitable, we added one random effects component for each of the scenarios (1) through (3), and for (4) where applicable, to the CCM and ESM and let the model selection procedure decide which one is best suited to describe the covariance observed in the data.

For the CCM, we thus effectively fitted the full model

$$\mathbb{E}[\boldsymbol{Y}] = h[\boldsymbol{S}\boldsymbol{B}\boldsymbol{1}_P^\top + \underset{\substack{(1)\\ \text{phylogenetic}}}{\tilde{\boldsymbol{S}}\boldsymbol{\Gamma}\boldsymbol{P}} + \underset{\substack{(2)\\ \text{spatial}}}{\boldsymbol{S}\boldsymbol{\Gamma}\tilde{\boldsymbol{P}}} + \underset{\substack{(3)\\ \text{spatio-}\\ \text{phylogenetic}}}{\tilde{\boldsymbol{S}}\boldsymbol{\Gamma}\tilde{\boldsymbol{P}}} + \underset{\substack{(4)\\ \text{independent}}}{\boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{P}}\ ] \tag{3}$$

and for the ESM we fitted

$$\mathbb{E}[\boldsymbol{Y}] = h[\boldsymbol{S}\boldsymbol{B}\boldsymbol{D}^\top + \underset{\substack{(1)\\ \text{phylogenetic}}}{\tilde{\boldsymbol{S}}\boldsymbol{\Gamma}\boldsymbol{X}} + \underset{\substack{(2)\\ \text{spatial}}}{\boldsymbol{S}\boldsymbol{\Gamma}\tilde{\boldsymbol{X}}} + \underset{\substack{(3)\\ \text{spatio-}\\ \text{phylogenetic}}}{\tilde{\boldsymbol{S}}\boldsymbol{\Gamma}\tilde{\boldsymbol{X}}} + \underset{\substack{(4)\\ \text{independent}}}{\boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{X}}\ ]. \tag{4}$$

For example, for the second random effects component, (2) spatial, we did not need to penalise $\boldsymbol{S}$ because we implicitly assumed a phylogenetic correlation matrix $\boldsymbol{\Sigma}_{\text{spp}} = \boldsymbol{I}_S$. For each random effect component, a variance estimate $\hat{\sigma}^2_\gamma$ is obtained. Ideally only one of these estimates is non-zero, and the correlation structure pertaining to its random effect component is thus *chosen*. When all estimates are zero, this indicates a complete lack of group-specific variance across species and plots in the data. Our aim was to identify the correlation structure most likely to be present in the data. After selecting a correlation structure, a second model was fitted, using only the random effect component of the chosen correlation structure.

If several variance estimates were positive, we thus needed to choose among these by model selection. Classical criteria for model selection, such as the AIC or BIC, are not adequate for model selection in the mixed models setting when the focus is on the choice of random effects (Vaida & Blanchard, 2005;

Braun et al., 2014). Vaida & Blanchard (2005) proposed a conditional AIC criterion for comparing linear mixed models with different random effects structures, based on inference on the conditional likelihood. The concept was extended to also apply to GLMMs by both Yu & Yau (2012) and Saefken et al. (2014). Another approach, suggested by Braun et al. (2014), uses mean cross-validated proper scores (see Gneiting & Raftery, 2007) to choose the model with the best predictive abilities in the GLMM setting. This method, which we utilised for the PSGLMMs, is suitable for choosing random as well as fixed effects.

Because the predictive distribution F_y of the data y is not analytically accessible, its first two central moments were obtained through a predictive cross-validation. For each plot p, we subsequently left one species s out and estimated its expectation μ_y and variance σ_y^2 based on the remaining species in plot p. For each plot-species datum, we calculated a proper score. For Gaussian and Poisson data, we applied the Dawid-Sebastiani score

$$S_{\mathrm{DS}}(F_y, y) = -\frac{1}{2}\left[\log \sigma_y^2 + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]$$

and for binary data, we applied the Brier score

$$S_{\mathrm{B}}(F_y, y) = -(\mu_y - y)^2.$$

For each model containing only one random effect component, the mean of all proper scores was calculated and compared with that of the other models before a final choice was made. For both scores, a larger score means better predictive abilities. We refer to Braun et al. (2014) for technical details.

The PSGLMMs may also be used to choose among several phylogenetic trees to decide which is best suited for the observed species data. If we assume that a spatio-phylogenetic signal is present, e.g. the CCM

$$\mathbb{E}[\boldsymbol{Y}] = h[\boldsymbol{S}\boldsymbol{B}\mathbf{1}_P^\top + \tilde{\boldsymbol{S}}^{(A)}\boldsymbol{\Gamma}\tilde{\boldsymbol{P}} + \tilde{\boldsymbol{S}}^{(B)}\boldsymbol{\Gamma}\tilde{\boldsymbol{P}}]$$

performs this test for two different phylogenetic correlation matrices Σ^A_{spp} and Σ^B_{spp}, derived from phylogenetic trees A and B. $\tilde{\boldsymbol{S}}^{(A)}$ and $\tilde{\boldsymbol{S}}^{(B)}$ are the species matrices penalised with the Cholesky factor corresponding to each phylogenetic correlation matrix.

3 Simulation Study

We assessed the performance of the PSGLMMs in simulation studies. For both the CCM and ESM, the three settings (a) *phylogenetic*, (b) *spatial* and (3) *spatio-phylogenetic* were simulated for each response type (Gaussian, binary, and Poisson). For all cases, responses were simulated for $S = 4$

species (spp) in $P = 100$ plots aligned on a 10×10 grid. The phylogenetic correlation matrix was set to

$$\mathbf{\Sigma}_{\text{spp}} = \begin{bmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where $\rho \in [0.02, 0.98]$ is the correlation between species 1 and 2. The responses of all other pairs of species were uncorrelated. The spatial correlation matrix $\mathbf{\Sigma}_{\text{plot}}$ was calculated using the Matérn correlation function, as given by Rue & Held (2005):

$$\frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{d}{\phi}\right)^{\kappa} K_{\kappa} \left(\frac{d}{\phi}\right),$$

with κ and ϕ both set to 1, and where d is the Euclidean distance between the centroids of any two plots and K_{κ} is the Bessel function. All plots were unit squares. To reduce computation time, all correlations < 0.10 were set to zero. This implied that the nearest 20 neighbours were used with a spatial correlation ranging from 0.11 to 0.37, independent of the phylogenetic correlation ρ and constant for all simulation runs.

For the ESM, we assumed that all base responses of species to the envi-

ronmental factor were zero. The first terms of the linear predictor matrices $\boldsymbol{\eta}_{\text{CCM}} = \boldsymbol{S}\boldsymbol{B}\mathbf{1}_P^\top + \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{P}$ and $\boldsymbol{\eta}_{\text{ESM}} = \boldsymbol{S}\boldsymbol{B}\mathbf{1}_P^\top + \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{X}$ were thus identical to $\boldsymbol{B} = \boldsymbol{\beta}_{\text{spp}}$, as given in Table 3. $\boldsymbol{S}$ and $\boldsymbol{P}$ were identity matrices of size S and P respectively and $\boldsymbol{X}$ was a 10×10 array containing the values in each plot of the environmental factor $\boldsymbol{x}$. These were drawn from a P-dimensional normal distribution with mean vector $\boldsymbol{\mu}_x = \begin{bmatrix} 1 & 1 + 1/99 & \dots & 2 \end{bmatrix}^\top$ and covariance matrix $\boldsymbol{\Sigma}_x = 0.25 \cdot \boldsymbol{\Sigma}_{\text{plot}}$. The resulting values $\boldsymbol{x}$ ranged from 0.27 to 3.05 and exhibited a slight spatial pattern, with lower values in the lower left part and higher values in the upper right part of the array. The random effects $\boldsymbol{\Gamma}$ were simulated in accordance with Table 1, and the responses obtained were as follows:

- Gaussian: $\text{vec}(\boldsymbol{Y}) \sim \mathcal{N}_{SP}(\text{vec}(\boldsymbol{\eta}), \sigma_\varepsilon^2 \boldsymbol{I}_{SP})$ with residual variance $\sigma_\varepsilon^2 = 0.01$
- Binary: $\boldsymbol{Y}_{sp} \sim \mathcal{B}(1, h(\boldsymbol{\eta}_{sp}))$, where $h(\boldsymbol{\eta}_{sp}) = \mathbb{P}[\boldsymbol{Y}_{sp} = 1]$ and the response function is $h(\boldsymbol{\eta}) = \exp\{\boldsymbol{\eta}\}/(1 + \exp\{\boldsymbol{\eta}\})$
- Poisson: $\boldsymbol{Y}_{sp} \sim \mathcal{P}(h(\boldsymbol{\eta}_{sp}))$, where $h(\boldsymbol{\eta}_{sp}) = \mathbb{E}[\boldsymbol{Y}_{sp}]$ and $h(\boldsymbol{\eta}) = \exp\{\boldsymbol{\eta}\}$

with $s = 1, \dots, S$ and $p = 1, \dots, P$.

[Table 1 about here.]

In the phylogenetic setting (a), the correlation of the simulated responses

thus increased with ρ as it depended on the phylogenetic correlation matrix $\mathbf{\Sigma}_{\text{spp}}$. In the spatial setting (b), the correlation of the responses was constant for all ρ, as it only depended on the spatial correlation matrix $\mathbf{\Sigma}_{\text{plot}}$. In the spatio-phylogenetic setting (c), the correlation rose with ρ, but was also dependent on the spatial correlation matrix.

The phylogenetic and spatio-phylogenetic random effect components (1) and (3) in equations 3 and 4 also varied with ρ as the penalised species matrix $\tilde{\boldsymbol{S}}$ depended on the Cholesky factor of $\mathbf{\Sigma}_{\text{spp}}$, i.e. the same phylogenetic correlation ρ used for generating the data also defined Σ_{spp} in the model. The spatial and independent random effect components (2) and (4) were fixed for all ρ. For each simulation setting, the *correct* random effect component chosen by the model was thus the one carrying the same name as the setting. For each response type and setting, 100 data sets were simulated. The resulting parameter estimates are found in Figure 1 and 3 and 4 in the online supplement for the CCMs and in Figures 5–7 in the online supplement for the ESMs. The results for each simulation setting are given in separate columns, with one plot for each random effect component modelled. For low correlations ρ, $\mathbf{\Sigma}_{\text{spp}}$ was similar to the identity matrix $\boldsymbol{I}_S$. In the phylogenetic simulation setting (a), the phylogenetic and independent random effect components were thus also similar for low correlations, and we expected the estimated variance to be split between them. This was also the case for the

spatial and spatio-phylogenetic random effect components in the spatial and spatio-phylogenetic settings (b) and (c). For increasing correlations, we expected the variance estimate pertaining to the correct component, i.e. the one used to simulate the data in that setting, to level off around the true parameter, while all other variance estimates were expected to tend towards zero.

Not all simulated runs could be evaluated. Especially for Gaussian responses (between 16% and 19% of all runs were discarded for the CCM and between 25% and 29% for the ESM) and for Poisson responses for the ESM (between 2% and 5% of all runs were discarded), a considerable number were discarded (see Table 2). For Gaussian responses, all discarded runs were due to the parameter estimation (in `lme4`) not converging due to nonidentifiable variance and residual variance parameters, see Section 2.3. For Poisson responses for the ESM, most discarded runs were due to numerical issues pertaining to the predictive cross-validation, see Braun et al. (2014).

[Table 2 about here.]

3.1 Results CCM

The variance parameter estimates for the CCM for Gaussian responses in Figure 1 all clearly followed the expected trends. For low values of ρ in the

phylogenetic setting, the estimated phylogenetic variance $\hat{\sigma}^2_{\text{spp}}$ and the residual variance $\hat{\sigma}_\epsilon$ were close to one, although the corresponding true parameters were $\sigma^2_{\text{spp}} = 2$ and $\sigma_\epsilon = 0.01$ as a consequence of near nonidentifiability. In the spatial setting (b), the results depended on ρ only indirectly via the phylogenetic term in the model. Low values of ρ rendered parameter estimation more difficult and were associated with a higher variability of $\hat{\sigma}^2_{\text{plot}}$. It is plausible that the estimates in the spatio-phylogenetic setting (c) were less precise than the estimates in the phylogenetic (a) and spatial (b) settings, as the variance parameter estimated was made up of two separate terms ($\sigma^2_{\text{spp}}\sigma^2_{\text{plot}} = 2 \cdot 2 = 4$). For all cases, the parameter estimates levelled off at values quite close to the true parameters (see Table 3).

[Table 3 about here.]

[Figure 1 about here.]

For binary responses, the expected trends were less apparent (see Figure 3 in online supplement). A positive random effect variance parameter estimate in the phylogenetic setting (a) was not notable until the correlation ρ reached a medium strength, but the estimate rose steadily. This positive trend was also clear for the spatio-phylogenetic setting (c); in the spatial setting (b), the estimations were split between the spatial and spatio-phylogenetic components for all correlations. This was caused by the relatively weak spatial

correlation structure Σ_{plot} used. All variance estimates were highly variable. The results for Poisson responses (see Figure 4 in online supplement) were more or less as in the case of Gaussian responses, except with considerably more variability to the parameter estimates. The parameter estimates also tended more slowly towards the true parameters than estimates for Gaussian responses.

Figure 2 depicts the selection rates of the correct random effect components, based on the predictive cross-validation, for each response type and simulation setting for the CCM. The red line illustrates the selection rate of the correct random effect component; for example, in the phylogenetic setting (a), it indicates the selection rate of the random effect component phylogenetic. The blue line illustrates the rate with which any of the two components in the phylogenetic or spatio-phylogenetic setting were selected, i.e. at which rate any phylogenetic signal was detected. In the spatial setting (b), the latter rate is thus false positive and should ideally be low.

[Figure 2 about here.]

All selection rates for Gaussian responses reached a very high level for relatively low correlations ρ. For the spatio-phylogenetic setting (c), however, a severe drop in the selection rate for high correlations contradicted the tendency of the variance parameter estimates in Figure 1. As can be seen from

the blue curve, this was due to the phylogenetic component being chosen instead.

The selection rates for binary responses also reached a relatively high level for the phylogenetic and spatio-phylogenetic settings (a) and (c). For the spatial setting (b), however, the inability of the model to choose between the spatial and spatio-phylogenetic components became even clearer, with a selection rate constantly at around 0.50. The results indicate that it is in fact rather difficult for the model and model selection procedure to decide which correlation structure (spatio-phylogenetic or spatial) fits the data best, when in fact only a moderately large spatial correlation is present in binary or count data as was the case in our simulation study.

For Poisson responses as well, the selection rate of the correct component in the spatial setting (b) was low. This trend was due to the phylogenetic and spatio-phylogenetic components being chosen for Gaussian and binary responses, whereas for Poisson responses, the independent component (4) was chosen most often. In the spatio-phylogenetic setting (c), only a medium selection rate of around 0.50 was reached for the correct component, although the selection rate of either of the two components, phylogenetic and spatio-phylogenetic, was relatively high for medium to strong correlations.

3.2 Results ESM

For the ESMs, the trends of the parameter estimates were more precise and clearer than for the CCM for all response types, but otherwise did not differ noticeably (see Figures 5–7 in the online supplement).

The ESM selection rates (Figure 3) differed from those of the CCM. While the selection rate in the phylogenetic setting (a) was somewhat higher for binary responses, it was considerably lower for Poisson responses for medium to high correlations. On the other hand, in the spatio-phylogenetic setting (c), the selection rate was much better for Poisson responses and somewhat worse for Gaussian responses. The selection rate for Gaussian responses in the spatial setting (b) was also much lower than for the CCM. However, the false positive rate was approximately the same.

[Figure 3 about here.]

3.3 Comparison to PGLMMs

To compare the performance of the original implementation of the phylogenetic generalised linear mixed models (PGLMMs) of Ives & Helmus (2011) with our `lme4`-based implementation of the PSGLMMs, we fitted both models to 100 simulated binary occurrence data sets and compared the results. As the models in the special case of spatially uncorrelated data are the same,

differences can be attributed to different software implementations used. The estimation technique applied for the PGLMMs by Ives & Helmus' (2011) in the R package `picante` (Kembel et al., 2010) is a combination of penalised quasi-likelihood (PQL) and REML estimation. For the PSGLMMs, which make use of the R package `lme4` (Bates et al., 2013), the random effect variance parameters are estimated via Laplace approximation.

The data were simulated using the function `pglmm.sim()` from `picante`; the corresponding simulation model is described in Ives & Helmus (2011). For each simulation run, the occurrence of $S = 16$ species was simulated in $P = 30$ plots. During simulation, plots with fewer than two species present were discarded. The actual number of plots may thus be smaller and may differ between the simulation runs. The phylogenetic relationship of the species was characterised through a balanced phylogenetic tree, from which a phylogenetic correlation matrix $\boldsymbol{\Sigma}_{\text{spp}}$ was derived. The species-specific random effects were simulated using a variance estimate $\sigma^2_{\text{spp}} = 1$. A random effect accounting for differences in the number of species in each plot was also added, with a small but positive variance parameter σ^2_{plot}.

In array form, the model was set up as

$$\mathbb{E}[\boldsymbol{Y}] = h[\boldsymbol{S}\boldsymbol{B}\boldsymbol{1}_P^\top + \tilde{\boldsymbol{S}}\boldsymbol{\Gamma}\boldsymbol{P} + \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{P}],$$

with random effects $\boldsymbol{\Gamma} \sim \mathcal{N}_{SP}(0, \sigma^2_\gamma \boldsymbol{I}_P \otimes \boldsymbol{I}_S)$, fixed effects $\boldsymbol{B} = \boldsymbol{\beta}_{\text{spp}}$ and $\tilde{\boldsymbol{S}} = \boldsymbol{S}\boldsymbol{R}_{\Sigma_{\text{spp}}}$. $\boldsymbol{S}$ and $\boldsymbol{P}$ are identity matrices of size S and P. The first random effect component models the random effects of the species, while accounting for their phylogenetic relatedness. The second component models an unstructured random effect in the plots.

In their original model, Ives & Helmus (2011) assume a variance-covariance structure $\sigma^2_{\text{plot}} \boldsymbol{I}_P \otimes \boldsymbol{J}_S$ for the random effect in the plots, where $\boldsymbol{J}_S$ is an $(S \times S)$-matrix of ones. Because this matrix is not positive semi-definite it cannot be used for our PSGLMMs. In order to allow for the results from the two models to be compared we instead utilised the variance-covariance structure $\sigma^2_{\text{plot}} \boldsymbol{I}_P \otimes \boldsymbol{I}_S$ for both models.

The variance parameters σ^2_{spp} and σ^2_{plot} were estimated for the PGLMM using the function `pglmm.fit()` from `picante`. For the PSGLMM, we used the code given in the online supplements, which made use of `glmer()` from `lme4`. In addition to obtaining the random effect variance estimates with both methods, we also calculated the REML log-likelihood and considered the fixed effects obtained with the two different methods. As the estimation technique was the only difference between the fitted models, we expected their results, presented in Figures 1 and 2 in the online supplement, to be similar. Note that 28 runs were discarded, either because the PGLMM (5) or the PSGLMM (18) estimation algorithm failed to converge or because of issues

of quasi-complete separation (5).

The log-likelihood was practically identical for both models in all considered runs. The variance estimates $\hat{\sigma}^2_{\text{spp}}$ were in general higher and varied more for the PSGLMM than for the PGLMM, but the estimates of the two models were highly correlated. The median estimate was 0.578 for the PGLMM and 0.788 for PSGLMM. Regarding the estimates $\hat{\sigma}^2_{\text{plot}}$ for the PGLMM and PSGLMM, we observed no coherence other than most estimates being close to zero. The fixed effect parameter estimates were also highly correlated between the two models, although they were generally somewhat lower for the PSGLMM.

3.4 Summary of Simulation Results

In general, PSGLMMs based on their implementation in the package `lme4` were able to estimate the fixed and variance parameters and the suggested model selection procedure performed well in selecting the correct model in spatio-phylogenetically correlated Gaussian, binary, and Poisson data.

Exceptions were observed in the following situations: For Gaussian responses with a low phylogenetic correlation, the variance parameters pertaining to phylogenetic and spatio-phylogenetic model terms and the residual variance were not identifiable resulting in biased estimates with a large variability.

Although we did not alter the spatial correlation here, the same would apply to low spatial correlations in a data set. Therefore, one cannot expect the procedure to work well for low phylogenetic or spatial correlations. For higher correlations ($\rho > 0.1$ in our simulations), the model performed very well in all three scenarios and the model selection procedure led to the correct model with high probability, although it seemed to be hard to differentiate between phylogenetic and spatio-phylogenetic terms for very large correlations.

For binary responses, the variability of the estimated variance parameters was much higher, due to the lower information contained in binary data. As a consequence, the selection frequency of the correct model was not quite as high as in the Gaussian case but still acceptable. Note that the low selection frequency in the spatial setting (b) was due to the relatively low spatial correlation used in our simulations. For more strongly correlated observations, the corresponding selection frequency would be much higher. The same applies to the Poisson case.

Our simulation results for the phylogenetic setting (a) carry over to the traditional PGLMM model because the two software implementations (based on the `nlme` and `lme4` packages) led to practically identical results.

4 Application to Bird Abundance Data

In the following section, the PSGLMMs were applied to a data set of bird counts to test for phylogenetic signals. The data consist of observational counts of 47 bird species belonging to the orders *Piciformes* and *Passeriformes*, made along four transects through the Bavarian Forest National Park in Germany (see Figure 8 in the online supplement). Counts ranging from 0 to 37 were made on a total of 371 plots. With 74% of the data being zeros, the empirical distribution was strongly positively skewed.

We chose a subgroup of six closely related species consisting of the passerine (*Passeriformes*) species *Poecile montanus*, *Periparus ater*, *Lophophanes cristatus*, *Poecile palustris*, *Parus major* and *Cyanistes caeruleus*, with a mean phylogenetic correlation of 0.808 between the species. The spatial correlation matrix was calculated on the basis of the Euclidean distances between the plots using the Matérn correlation function. To reduce computational time, spatial correlations < 0.05 were set to zero. For the ESM, we considered five environmental factors: elevation, coverage (in %) of the middle forest layer by broadleaved trees (in the following referred to as *broadleaved trees*) and by rejuvenating broadleaved trees (*rejuvenation*), the maximum trunk diameter (in cm) at breast height (DBH) and the total number of tree holes within an area of $0.1ha$ (*tree holes*).

4.1 Results

In addition to fitting a CCM, we fitted one ESM for each environmental factor. To all models, we added the six random effect components listed in Table 4. The component *spatial* covers the case where the variance is of a purely spatial nature – independent of the phylogenetic relatedness of the birds. To cover the cases where the variance is dependent on only phylogeny or on phylogeny and space, we added the components *phylogenetic* and *spatio-phylogenetic* twice – once for an attraction tendency and once for a repulsion tendency (see Ives & Helmus, 2011). For an attraction tendency, we assumed that closely related species are more likely to co-occur than more distantly related species. We utilised the phylogenetic correlation matrix to incorporate this case. For a repulsion tendency, we assumed that closely related species are less likely to co-occur than more distantly related species. Following Ives & Helmus (2011) we utilised the inverse of the phylogenetic correlation matrix, $\mathbf{\Sigma}_{\text{spp}}^{-1}$, for this case. Lastly, we also added the component *independent* to cover the possibility of an observed variance in the sensitivity of the species to the environmental factors to be independent of space and phylogeny.

[Table 4 about here.]

The choice of a random effect component by the PSGLMM thus allowed us to decide whether or not a phylogenetic signal was detected in the sensitivity of the species to an environmental factor, and if so, whether closely related species have a tendency to attract or repel each other. Fixed intercepts were added for each species to account for their overall abundance. To account for the different base sensitivity of the species to the environmental factors, fixed slopes in the environmental factors were also added for each species. For the CCM, we expected either the phylogenetic or the spatio-phylogenetic component with an attraction tendency to be chosen, and this was the case: a spatio-phylogenetic community structure in which closely related species are more likely to co-occur (attraction) was chosen by the CCM with an estimated random effect variance of $\hat{\sigma}^2_\gamma = 1.860$. Because of the spatial patterns of the land-cover types, we expected either the spatial or one of the spatio-phylogenetic random effect components to be chosen for the ESMs. We further expected an attraction tendency as these birds species are not known to compete for resources.

A spatio-phylogenetic signal was detected in the sensitivity of the species to tree holes, broadleaved trees, and rejuvenation (see Table 5). For the latter two environmental factors, the species exhibited an attraction tendency, i.e., the closer the species were related, the more likely they were to share similar values of these environmental factors. For tree holes, the species exhibited

a repulsion tendency, with more distantly related species being more likely to share similar values. For the environmental factors DBH and elevation, a purely spatial correlation structure was chosen. The species were thus more likely to share similar values when spatially close to each other, regardless of their phylogenetic relatedness.

[Table 5 about here.]

5 Conclusion

Numerous studies have already provided convincing evidence that ecological communities exhibit phylogenetic structure at both small and large spatial scales (e.g. Graves & Gotelli, 1993; Webb et al., 2002; Lovette & Hochachka, 2006; Helmus et al., 2007; Pillar & Duarte, 2010; Riedinger et al., 2013). Most of these authors used one or several of the available metrics (e.g. mean phylogenetic distance between species or individuals) to summarise the phylogenetic structure as well as various null-models to explore whether the selected metric deviates from the expectation of a random distribution of species. But as noted by Ives & Helmus (2011), such metrics summarise the complexities of the community structure in a single number and therefore only allow for tests of basic hypotheses. Nevertheless, the use of multiple metrics in a study may increase our confidence in the analyses and interpre-

tations as long as all metrics indicate the same pattern. In the case where the metrics signal different patterns, we have the problem either to dismiss the pattern or to make an often arbitrary selection of a metric we want to rely on for further interpretations. To overcome these limitations, Ives & Helmus (2011) have implemented generalised linear mixed models for phylogenetic community analyses. They have shown that such mixed models have a higher statistical power in detecting phylogenetic signals than approaches that rely on metrics and null-models.

Considering PGLMMs as special cases of GLAMs has three merits: (1) the extension to more complex models, such as one allowing a spatial correlation between observations, is possible; (2) computationally attractive properties of GLAMs will carry over to P(S)GLMMs; and (3) novel model selection procedures can be applied. Although Currie et al. (2006) present a prototype implementation of generalised linear mixed models, no production-ready code is currently available. We therefore used the `lme4` package to implement two of the models proposed by Ives & Helmus (2011): one model used the qualitative or quantitative composition of the community to test for a phylogenetic structure in a community matrix (Model I of Ives & Helmus, 2011), and the other model tested for a phylogenetic signal in the sensitivity of the species to environmental factors. Although the approaches are not identical, a comparison of the model fits using our PSGLMM and the implementation

of Ives & Helmus' (2011) PGLMM using simulated data indicated that the two approaches indeed perform on par in this special situation. Hence, our empirical findings in the scenario phylogenetic (a) carry over to the corresponding implementation of the phylogenetic models described by Ives & Helmus (2011). However, the novel model formulation presented here allows the spatial correlation structure between plots to be taken into account. In addition, selection between various correlation structures was also introduced for PSGLMMs.

PSGLMMs are very complex models relying on a number of rather strict assumptions. All limitations of the generalised linear mixed model framework apply, such as the conditional distribution of the response coming from an exponential family, conditional independence of the observations, nuisance parameters being independent of explanatory variables, and independent normal random effects. The latter assumption might be hard to justify since neither the species- nor plot-specific random effects can be assumed to by symmetric in general. The ability to let the model select one of several phylogenetic correlation structures relaxes strict assumptions on the phylogenetic signal. In principle, this would also be possible for the spatial correlation, however, the Matérn family is one of the most flexible approaches and a standard choice in spatial statistics (Fahrmeir et al., 2013).

Our simulation studies showed that PSGLMMs are generally capable of se-

lecting the correct random effect components and yield precise estimates of the variance parameters. Although the selection rate of the correct random effect component was somewhat low for low correlations, the models were effective in detecting whether a phylogenetic signal (spatial or not) was present. However, both the CCM and the ESM exhibited a quite high false-positive rate for binary and Poisson responses. For binary occurrence data, it is important that rare or pervasive species are not included when fitting the models as these often lead to issues concerning quasi-complete separation, which leads to estimates of very large positive or negative fixed effects. For Poisson responses, especially rare species are problematic. A technical consequence of the large negative fixed effects is that they render the predictive cross-validation infeasible, as weights pertaining to these species are set to zero and later used as denominators. Overall, one cannot expect PSGLMMs to identify the correct correlation structure (phylogenetic, spatial, or spatio-phylogenetic) for low correlations. When the correlations can be assumed to be rather large, PSGLMMs and the model selection procedure have the potential to select the correlation structure that comes closest to reality. Improved optimisers able to deal with the near-nonidentifiability of the variance parameters in the Gaussian case may offer improvements in the situation of low phylogenetic or spatial correlations.

The computational restrictions in our study considering only six species can

be relaxed by utilising the array structure of the data. This would reduce computational time drastically and would allow us to fit the models to large community matrices in situations where convergence of the mixed model estimation is unproblematic. Because of the small number of species used, our example should be considered as an illustration. Nevertheless, the application to the bird count data in Bavaria revealed spatio-phylogenetic signals both in the community structure and in the species' sensitivity to several environmental factors. Thus, the application of PSGLMMs allowed the detection of sophisticated spatial-phylogenetic patterns.

Acknowledgements

We thank Karen A. Brune for improving the language.

Data Accessibility

Data and computer code for data analysis and simulation experiments are contained in the R add-on package **TH.data**, version 1.0-5 or higher, available from `http://CRAN.R-project.org/package=TH.data`.

References

Bates, D.; Maechler, M.; Bolker, B. & Walker, S. (2013) *lme4: Linear mixed-effects models using Eigen and S4*, R package version 1.0-0.

Blomberg, S.P. & Garland, T. (2002) Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods, *Journal of Evolutionary Biology*, 15(6), 899–910.

Blomberg, S.P.; Garland, T. & Ives, A.R. (2003) Testing for phylogenetic signal in comparative data: Behavioral traits are more labile, *Evolution*, 57(4), 717–745.

Braun, J.; Sabanés Bové, D. & Held, L. (2014) Choice of generalized linear mixed models using predictive crossvalidation, *Computational Statistics & Data Analysis*, 75, 190–202.

Currie, I.D.; Durban, M. & Eilers, P.H.C. (2006) Generalized linear array models with applications to multidimensional smoothing, *Journal of the Royal Statistical Society Series B*, 68(2), 259–280.

Fahrmeir, L.; Kneib, T.; Lang, S. & Marx, B. (2013) *Regression: Models, Methods and Applications*, Springer.

Felsenstein, J. (1985) Phylogenies and the comparative method, *The American Naturalist*, 125(1), 1–15.

Gneiting, T. & Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association*, 102, 359–378.

Graves, G.R. & Gotelli, N.J. (1993) Assembly of avian mixed-species flocks in Amazonia, *Proceedings of the National Academy of Sciences of the United States of America*, 90(4), 1388–1391.

Hansen, T.F. & Orzack, S.H. (2005) Assessing current adaptation and phylogenetic inertia as explanations of trait evolution: The need for controlled comparisons, *Evolution*, 59(10), 2063–2072.

Helmus, M.R.; Savage, K.; Diebel, M.W.; Maxted, J.T. & Ives, A.R. (2007) Separating the determinants of phylogenetic community structure, *Ecology Letters*, 10(10), 917–925.

Ives, A.R. & Helmus, M.R. (2011) Generalized linear mixed models for phylogenetic analyses of community structure, *Ecological Monographs*, 81(3), 511–525.

Kembel, S.; Cowan, P.; Helmus, M.; Cornwell, W.; Morlon, H.; Ackerly, D.;

720 Blomberg, S. & Webb, C. (2010) Picante: R tools for integrating phyloge-
721 nies and ecology, *Bioinformatics*, 26, 1463–1464.

722 Losos, J.B. (2008) Phylogenetic niche conservatism, phylogenetic signal and
723 the relationship between phylogenetic relatedness and ecological similarity
724 among species, *Ecology Letters*, 11(10), 995–1003.

725 Lovette, I.J. & Hochachka, W.M. (2006) Simultaneous effects of phyloge-
726 netic niche conservatism and competition on avian community structure,
727 *Ecology*, 87(7), 14–28.

728 Pillar, V.D. & Duarte, L.d.S. (2010) A framework for metacommunity anal-
729 ysis of phylogenetic structure, *Ecology Letters*, 13(5), 587–596.

730 R Core Team (2013) *R: A language and environment for statistical computing*,
731 R Foundation for Statistical Computing, Vienna, Austria.

732 Revell, L.J.; Harmon, L.J. & Collar, D.C. (2008) Phylogenetic signal, evolu-
733 tionary process, and rate, *Systematic Biology*, 57(4), 591–601.

734 Riedinger, V.; Müller, J.; Stadler, J.; Ulrich, W. & Brandl, R. (2013) Assem-
735 blages of bats are phylogenetically clustered on a regional scale, *Basic and*
736 *Applied Ecology*, 14(1), 74–80.

737 Rodríguez-Álvarez, M.X.; Lee, D.J.; Kneib, T.; Durbán, M. & Eilers, P.

(2014) Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm, *Statistics and Computing*, in press.

Rue, H. & Held, L. (2005) *Gaussian Markov random fields: Theory and applications*, vol. 104 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London.

Saefken, B.; Kneib, T.; van Waveren, C.S. & Greven, S. (2014) A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models, *Electronic Journal of Statistics*, 8, 201–225.

Stephens, P.R. & Wiens, J.J. (2009) Bridging the gap between community ecology and historical biogeography: niche conservatism and community structure in emydid turtles, *Molecular Ecology*, 18(22), 4664–4679.

Vaida, F. & Blanchard, S. (2005) Conditional Akaike information for mixed effects models, *Biometrika*, 92(2), 351–370.

Vamosi, S.M.; Heard, S.B.; Vamosi, J.C. & Webb, C.O. (2009) Emerging patterns in the comparative analysis of phylogenetic community structure, *Molecular Ecology*, 18(4), 572–592.

Webb, C.O.; Ackerly, D.D.; McPeek, M.A. & Donoghue, M.J. (2002) Phylogenies and community ecology, *Annual Review of Ecology and Systematics*, 33, 475–505.

Yu, D. & Yau, K.K.W. (2012) Conditional Akaike information criterion for generalized linear mixed models., *Computational Statistics & Data Analysis*, 56(3), 629–644.

List of Tables

1 Distributions used to simulate the random effects for the different correlation structures in each setting (a) through (c). . . 46

2 The number of discarded runs in the simulation study (from a total of $4,900$ runs per response type, setting and model). . 47

3 Fixed effects parameters $\boldsymbol{\beta}_{\text{spp}}$ for each of the four species and phylogenetic and spatial random effect variance parameters, σ^2_{spp} and σ^2_{spp} respectively, used for each response type in the simulation study of the PSGLMMs. 48

4 Variance-covariance matrices for each random effect component used to model the bird abundance data. For a *repulsion* tendency, where closely related species are less likely to co-occur, contrary to the attraction tendency, the inverse of the phylogenetic correlation matrix was used. 49

5 For each of five environmental factors, a separate ESM was fitted. With the spatial correlation structure chosen for DBH and elevation, we concluded that no phylogenetic signal is present in the response of the species to the factors. For the other factors, a spatio-phylogenetic signal was detected with a repulsion tendency for tree holes and an attraction tendency for broadleaved trees and rejuvenation. $\hat{\sigma}^2_\gamma$ is the random effect variance estimate for the correlation structure chosen using a predictive cross-validation. 50

Correlation structure	$\text{vec}(\boldsymbol{\Gamma}) \sim$	Simulation
(1) phylogenetic	$\mathcal{N}_{SP}(\mathbf{0}, \sigma^2_{\text{spp}} \boldsymbol{I}_P \otimes \boldsymbol{\Sigma}_{\text{spp}})$	Setting (a)
(2) spatial	$\mathcal{N}_{SP}(\mathbf{0}, \sigma^2_{\text{plot}} \boldsymbol{\Sigma}_{\text{plot}} \otimes \boldsymbol{I}_S)$	Setting (b)
(3) spatio-phylogenetic	$\mathcal{N}_{SP}(\mathbf{0}, \sigma^2_{\text{plot}} \sigma^2_{\text{spp}} \boldsymbol{\Sigma}_{\text{plot}} \otimes \boldsymbol{\Sigma}_{\text{spp}})$	Setting (c)
(4) independent	$\mathcal{N}_{SP}(\mathbf{0}, \sigma^2_{\text{spp}} \boldsymbol{I}_P \otimes \boldsymbol{I}_S)$	

Table 1: Distributions used to simulate the random effects for the different correlation structures in each setting (a) through (c).

	Response type	Setting (a) phylogenetic	Setting (b) spatial	Setting (c) spatio-phylogenetic
CCM	Gaussian	910	802	900
	binary	-	-	1
	Poisson	1	-	2
ESM	Gaussian	1,308	1,246	1,402
	binary	4	2	10
	Poisson	114	263	254

Table 2: The number of discarded runs in the simulation study (from a total of $4,900$ runs per response type, setting and model).

Response type	$\boldsymbol{\beta}_{\text{spp}}$	σ^2_{spp}	σ^2_{plot}
Gaussian	$[-2 \quad -2/3 \quad 2/3 \quad 2]^\top$	2	2
binary	$[-1 \quad -1/3 \quad 1/3 \quad 1]^\top$	1	1
Poisson	$[-1 \quad -1/3 \quad 1/3 \quad 1]^\top$	1	1

Table 3: Fixed effects parameters $\boldsymbol{\beta}_{\text{spp}}$ for each of the four species and phylogenetic and spatial random effect variance parameters, σ^2_{spp} and σ^2_{spp} respectively, used for each response type in the simulation study of the PSGLMMs.

Random effect component	Tendency	Variance-covariance matrix
spatial		$\sigma^2_\gamma \boldsymbol{\Sigma}_{\text{plot}} \otimes \boldsymbol{I}_S$
phylogenetic	attraction	$\sigma^2_\gamma \boldsymbol{I}_P \otimes \boldsymbol{\Sigma}_{\text{spp}}$
spatio-phylogenetic	attraction	$\sigma^2_\gamma \boldsymbol{\Sigma}_{\text{plot}} \otimes \boldsymbol{\Sigma}_{\text{spp}}$
phylogenetic	repulsion	$\sigma^2_\gamma \boldsymbol{I}_P \otimes \boldsymbol{\Sigma}^{-1}_{\text{spp}}$
spatio-phylogenetic	repulsion	$\sigma^2_\gamma \boldsymbol{\Sigma}_{\text{plot}} \otimes \boldsymbol{\Sigma}^{-1}_{\text{spp}}$
independent		$\sigma^2_\gamma \boldsymbol{I}_P \otimes \boldsymbol{I}_S$

Table 4: Variance-covariance matrices for each random effect component used to model the bird abundance data. For a *repulsion* tendency, where closely related species are less likely to co-occur, contrary to the attraction tendency, the inverse of the phylogenetic correlation matrix was used.

Environmental factor	Chosen correlation structure	Tendency	$\hat{\sigma}^2_\gamma$
elevation	spatial		< 0.001
broadleaved trees	spatio-phylogenetic	attraction	0.007
rejuvenation	spatio-phylogenetic	attraction	0.004
DBH	spatial		< 0.001
tree holes	spatio-phylogenetic	repulsion	0.030

Table 5: For each of five environmental factors, a separate ESM was fitted. With the spatial correlation structure chosen for DBH and elevation, we concluded that no phylogenetic signal is present in the response of the species to the factors. For the other factors, a spatio-phylogenetic signal was detected with a repulsion tendency for tree holes and an attraction tendency for broadleaved trees and rejuvenation. $\hat{\sigma}^2_\gamma$ is the random effect variance estimate for the correlation structure chosen using a predictive cross-validation.

List of Figures

1 Simulation study for the CCM with Gaussian responses: the distribution of the variance parameter estimates for each random effect component (rows) and setting (columns) subject to the phylogenetic correlation ρ between species 1 and 2. The true parameter to be estimated was $\hat{\sigma}^2_\gamma = 2$ in settings (a) and (b) and $\hat{\sigma}^2_\gamma = 2 \times 2 = 4$ in setting (c). Outliers are not drawn. 52

2 Selection rates for the CCM. The rate at which the correct random effect component (red line) was chosen by the CCM for each response type and setting (a) through (c) subject to the phylogenetic correlation ρ between species 1 and 2. The *correct component* is the component that was used to simulate the data in each setting (e.g. phylogenetic component in setting (a)). The blue line depicts the rate at which any of the random effect components (phylogenetic or spatio-phylogenetic) was chosen by the CCM and thus detects a (spatio-)phylogenetic signal. In setting (b), this rate thus acts as a false-positive rate and should be low. 53

3 Selection rates for the ESM. The rate at which the correct random effect component (red line) was chosen by the ESM for each response type and setting (a) through (c) subject to the phylogenetic correlation ρ between species 1 and 2. The *correct component* is the component that was used to simulate the data in each setting (e.g. phylogenetic component in setting (a)). The blue line depicts the rate at which any of the random effect components (phylogenetic or spatio-phylogenetic) was chosen by the CCM and thus detects a (spatio-)phylogenetic signal. In setting (b), this rate thus acts as a false-positive rate and should be low. 54

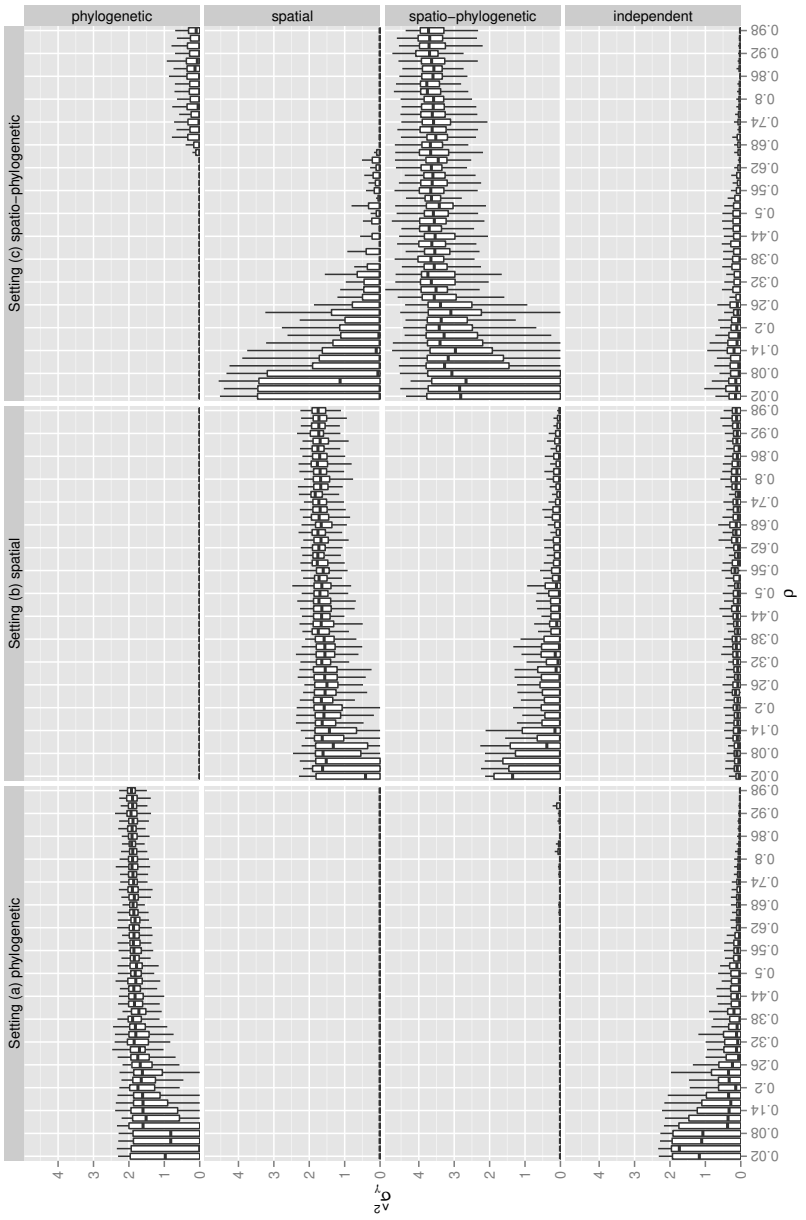

Figure 1: Simulation study for the CCM with Gaussian responses: the distribution of the variance parameter estimates for each random effect component (rows) and setting (columns) subject to the phylogenetic correlation ρ between species 1 and 2. The true parameter to be estimated was $\hat{\sigma}^2_\gamma = 2$ in settings (a) and (b) and $\hat{\sigma}^2_\gamma = 2 \times 2 = 4$ in setting (c). Outliers are not drawn.

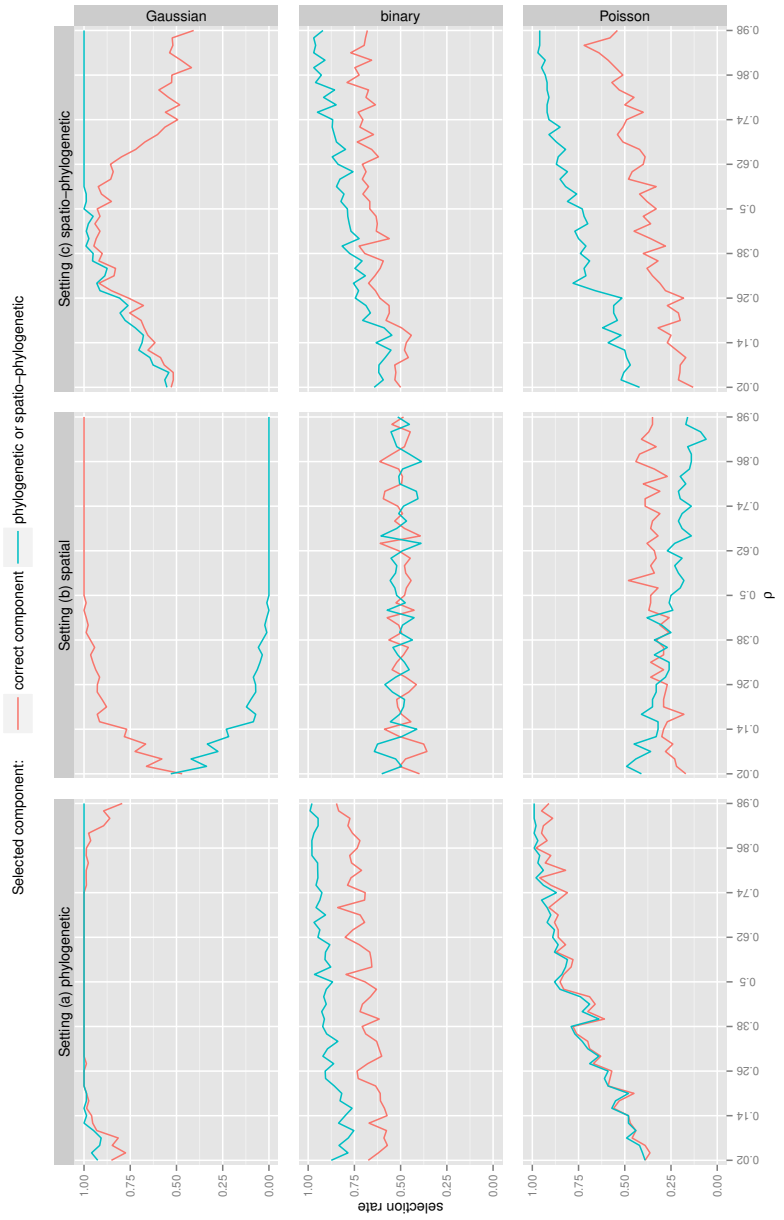

Figure 2: Selection rates for the CCM. The rate at which the correct random effect component (red line) was chosen by the CCM for each response type and setting (a) through (c) subject to the phylogenetic correlation ρ between species 1 and 2. The *correct component* is the component that was used to simulate the data in each setting (e.g. phylogenetic component in setting (a)). The blue line depicts the rate at which any of the random effect components (phylogenetic or spatio-phylogenetic) was chosen by the CCM and thus detects a (spatio-)phylogenetic signal. In setting (b), this rate thus acts as a false-positive rate and should be low.

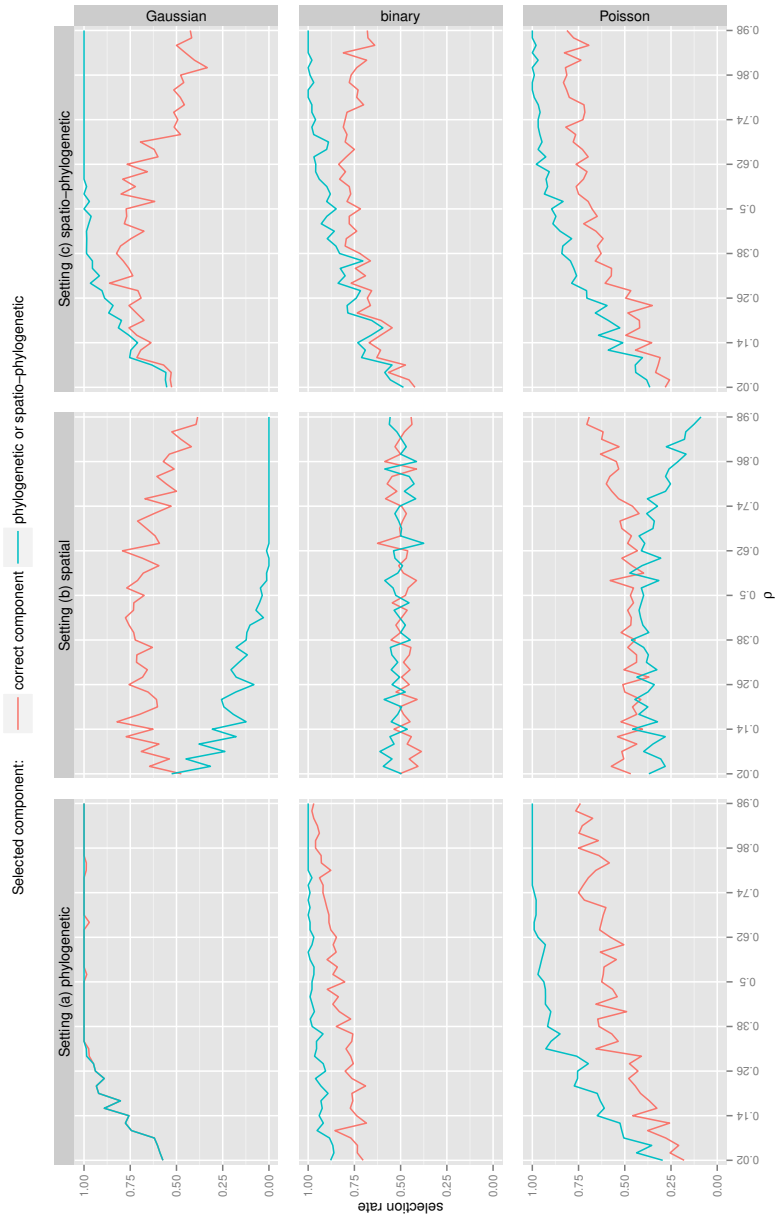

Figure 3: Selection rates for the ESM. The rate at which the correct random effect component (red line) was chosen by the ESM for each response type and setting (a) through (c) subject to the phylogenetic correlation ρ between species 1 and 2. The *correct component* is the component that was used to simulate the data in each setting (e.g. phylogenetic component in setting (a)). The blue line depicts the rate at which any of the random effect components (phylogenetic or spatio-phylogenetic) was chosen by the CCM and thus detects a (spatio-)phylogenetic signal. In setting (b), this rate thus acts as a false-positive rate and should be low.